No. F-14

会津大学

深層ニューラルネットワークの応答時間の短縮

教授　趙　強福

概要

〇深層ニューラルネットワーク(DMLP)は、左下の図のように、たくさんの「層」から構成される機械学習モデルである。従来、出力層は一つだけしかないので、任意の入力に対して、同じ計算量で出力を求める必要があり、平均応答時間が長い。
〇本技術で提案するDMLPは、複数の出力層を有する。任意の入力に対して、出力の「信頼度」が十分高くなった段階で計算を終了し、結果を出力することによって、平均応答時間を短縮することができる。データの「難易度」だけではなく、クラスの難易度を考慮に入れることもできる。
〇また、DMLPを訓練する際に、すべての層を一斉に求めるのではなく、層ごとに成長させていき、精度、信頼度などをもとに、ネットワークの成長を早期に終了し、層の数を最小限に抑えることもできる。
〇更に、各出力層は「理解可能な」学習モデルを使用することができる。例えば、決定木を利用することで、その層で抽出した「特徴」の中で、重要なものを自動的に選定できる。また、そのような特徴をもとに応答結果を「解釈」することもできる。

実用化の可能性

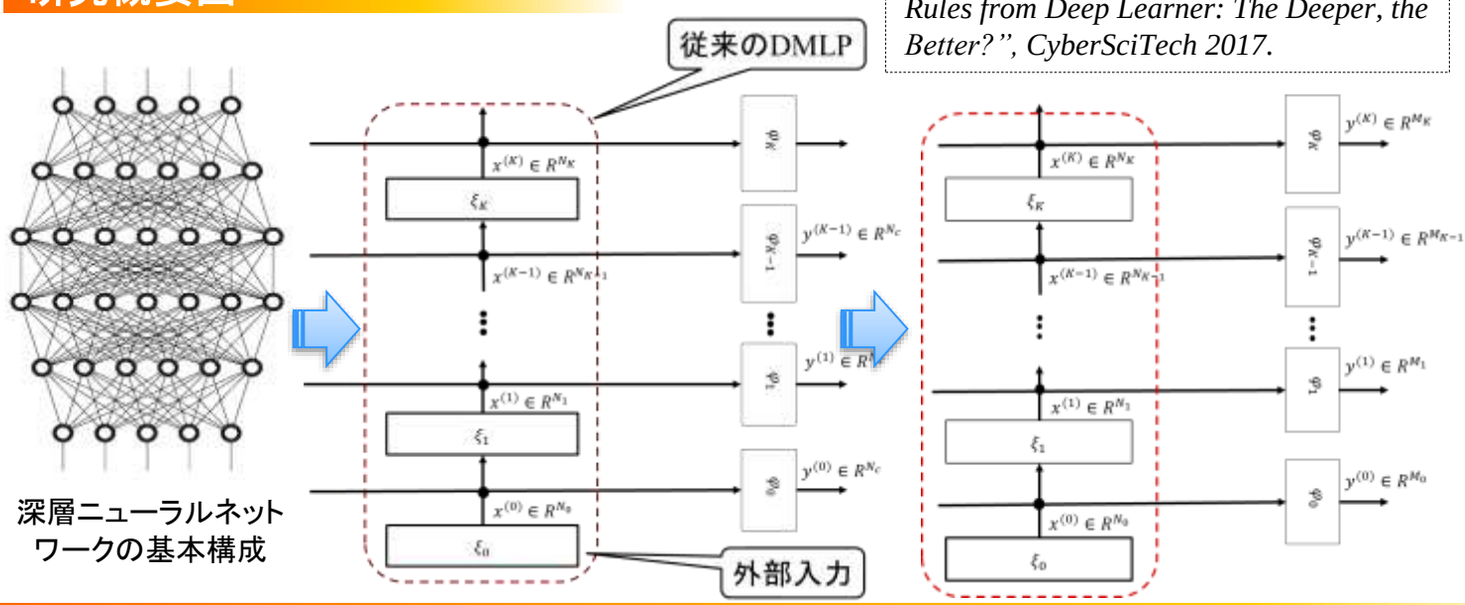
〇本技術は、DMLPの威力を発揮しつつ、応答に要する「平均時間」を短縮するために利用できる。例えば、NG製品の検出に応用する際、異常率が低い場合、正常品を早く、高い確率で判断できるようにモデルを設計すれば、検査時間を短縮することができる。また、自動運転の際には、危険があるかどうかを先に判断すれば、不必要な計算を省くことができる。要は、「大分類」、「細分類」、「細々分類」などを段階的に行うことによって、システムの実時間性を向上させることができる。

UBICからのメッセージ

〇昨今の深層ニューラルネットワークは、多層化により様々な問題を解決していますが、応答時間の増大や、内部処理のブラックボックス化が課題となっています。本技術は、これらに対する１つの解決策を与えるものです。今まで一律の処理時間を要していた問題を、難易度に応じて順次応答を返したり、各層で行われている処理の内容を外部から解釈できるようにすることも可能となります。

研究概要図

Furukawa and Zhao, "On Extraction of Rules from Deep Learner: The Deeper, the Better?", CyberSciTech 2017.



深層ニューラルネットワークの基本構成

よく現れるデータに、より速く反応する AI を作りましょう

関連発明：ニューラルネットワークの学習方法、コンピュータプログラム及びコンピュータ装置（特願2018-230323）